

Sparse Deep Stacking Network for Image Classification

Jun Li, Heyou Chang, Jian Yang

School of Computer Science and Technology

Nanjing University of Science and Technology, Nanjing, 219000, China.

junl.njust@gmail.com; changheyoun891001@126.com; csjyang@njust.edu.cn

Abstract

Sparse coding can learn good robust representation to noise and model more higher-order representation for image classification. However, the inference algorithm is computationally expensive even though the supervised signals are used to learn compact and discriminative dictionaries in sparse coding techniques. Luckily, a simplified neural network module (SNNM) has been proposed to directly learn the discriminative dictionaries for avoiding the expensive inference. But the SNNM module ignores the sparse representations. Therefore, we propose a sparse SNNM module by adding the mixed-norm regularization (l_1/l_2 norm). The sparse SNNM modules are further stacked to build a sparse deep stacking network (S-DSN). In the experiments, we evaluate S-DSN with four databases, including Extended YaleB, AR, 15 scene and Caltech101. Experimental results show that our model outperforms related classification methods with only a linear classifier. It is worth noting that we reach 98.8% recognition accuracy on 15 scene.

Introduction

It is well-known that sparse representations have a number of theoretical and practical advantages in computer vision and machine learning (Lee et al. 2007; Gregor and LeCun 2010; Yang et al. 2012). In particular, sparse coding techniques have led to promising results in image classification, e.g. face recognition and digit classification. Sparse coding, as a generative model, is a very important way to extract the sparse representations. However, sparse coding has the expensive inference algorithm and does not use the label of the training data. Although some researchers use the supervised signals to learn compact and discriminative dictionaries (Jiang, Lin, and Davis 2013; Zhuang et al. 2013; Huang et al. 2013), the expensive inference algorithm is still a problem. Since it is to train the dictionaries by using the labels, do we directly learn the discriminative dictionaries for avoiding the expensive inference?

Fortunately, a simplified neural network module (SNNM) (Deng and Yu 2011a) can directly train the discriminative dictionaries and fast calculate the representations. In the

SNNM, the input layer is non-linearly mapped to a hidden layer by using a projection matrix \mathbf{W} and a sigmoid activation function, and linearly mapped to an output layer by a matrix \mathbf{U} . Clearly, \mathbf{W} has discriminative ability because it is trained by minimizing the least squares error between the output vector and label vector. Moreover, SNNM can fast infer the hidden representation by only calculating a projection multiplication and a nonlinear transformation. Following a stacked scheme (Wolpert 1992), many SNNM modules are further stacked to build a Deep Stacking Network (DSN), which is previously named the Deep Convex Network (Deng and Yu 2011b). Recently, DSN has received increasing attentions due to its successful application in speech classification and information retrieval (Deng, Yu, and Platt 2012; Deng, He, and Gao 2013). Additionally, the DSN is attractive in that SNNM's the batch-mode nature offers a potential solution to the insurmountable problem of scalability in dealing with virtually unlimited amount of training data available nowadays (Deng and Yu 2013). Therefore, we extend DSN for image classification.

Despite DSN's success in speech classification, its framework also has several limitations. First, the conventional DSN only has used the sigmoid activation function for the nonlinear hidden layer (Deng, Yu, and Platt 2012). Although sigmoid has been widely used in the literature, it suffers from a number of drawbacks: for example the training can be slow, and with random initialization, the solution can stuck at a poor local solution that does not have good predictive performance (Glorot and Bengio 2010). In fact there are another two types of activation functions. The one is *hyperbolic tangent*, which has been applied to training deep neural networks. It suffers from the same problems as those of sigmoid functions. A more recent proposal is the *rectifier linear unit* (ReLU) (Nair and Hinton 2010). It is observed that this method is very useful for object recognition and often trains significantly faster (Glorot, Bordes, and Bengio 2011).

Second, sparse representations play a key role in image classification because they have the power to learn good robust features to noise, train gabor-like filters, and model more higher-order features (Ranzato et al. 2007; Lee, Ekanadham, and Ng 2008). Evidently, sparse representations have led to promising results in image classification (Jiang, Lin, and Davis 2013). Furthermore, there is considerable evidence that in brain the percentage of neurons active

is between 1 and 4% (Lennie 2003). It is reasonable to consider the sparse representations in SNNM modules. However, the conventional techniques for training SNNM completely ignores the sparse representations. Generally, they can be achieved by penalizing non-zero activation of hidden units (Ranzato, Boureau, and LeCun 2008) or a deviation of the expected activation of the hidden units (Lee, Ekanadham, and Ng 2008) in neural networks. Moreover, in neural networks the local dependencies between hidden units can make hidden units for better modeling observed data (Luo et al. 2011). But SNNM module restricted connections within hidden layer can not exhibit these dependencies. Fortunately, the hidden units without increasing the connections can be divided into non-overlapping groups for capturing the local dependencies among hidden units (Luo et al. 2011). The local dependencies can be implemented by using l_1/l_2 regularization upon the activation possibilities of hidden units in SNNM module.

In light of the above argument, this paper exploits a Sparse Deep Stacking Network (S-DSN) for image classification. S-DSN is obtained by stacking the sparse SNNM modules, which consider the two activation function: ReLU and sigmoid; and use the group sparse penalties (l_1/l_2 regularization) to penalize the hidden representations in SNNM module. Our S-DSN has many explicit advantages. First, compared with sparse coding technique (LC-KSVD (Jiang, Lin, and Davis 2013)), one-layer S-DSN can learn the projection dictionaries, which lead to a faster inference. Second, compared with DSN, S-DSN can extract sparse representations for learning good features in image classification. Last, S-DSN can retain the scalable structure of DSN. To conform the advantages of the S-DSN for image classification, extensive experiments have been performed on the four databases, including Extended YaleB, AR, 15 scene and Caltech101. Compared with multiple related methods, the experiments show that our model gets better classification results than other benchmark methods. In particular, we reach 98.8% recognition accuracy on 15 scene.

Deep Stacking Network

The DSN architecture is originally presented in the literature (Deng and Yu 2011b). Deng and Yu explore an original strategy for building deep networks, based on stacking layers of the basic SNNM modules, which take the simplified form of multilayer perceptron. We mathematically describe as follow:

Let the target vectors $\mathbf{t}_i = [t_{1i}, \dots, t_{ji}, \dots, t_{Ci}]^T$ be arranged to form the columns of $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_i, \dots, \mathbf{t}_N] \in R^{C \times N}$. Let the input data vectors $\mathbf{x}_i = [x_{1i}, \dots, x_{ji}, \dots, x_{Di}]^T$ be arranged to form the columns of $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_N] \in R^{D \times N}$. Formally, in the basic module, the lower-layer weight matrix, which is denoted by $\mathbf{W} \in R^{D \times L}$, connects the linear input layer and the nonlinear hidden layer. The upper-layer weight matrix, which is denoted by $\mathbf{U} \in R^{L \times C}$, connects the nonlinear hidden layer with the linear output layer. The outputs of upper-layer is $\mathbf{Y} = \mathbf{U}^T \mathbf{H}$, where $\mathbf{H} = \sigma(\mathbf{W}^T \mathbf{X}) \in R^{L \times N}$ is the hidden layer outputs and $\sigma(a) = 1/(1 + e^{-a})$

is the sigmoid activation function (Deng and Yu 2011a; Deng and Yu 2011b). The parameters \mathbf{U} and \mathbf{W} are learned to minimize the least squares objective:

$$\min_{\mathbf{U}, \mathbf{W}} f_{dsn} = \|\mathbf{U}^T \mathbf{H} - \mathbf{T}\|_F^2 + \alpha \|\mathbf{U}\|_F^2 \quad (1)$$

where α is a regularization parameter of upper-layer weight matrix \mathbf{U} .

Clearly, \mathbf{U} has a closed-form solution:

$$\mathbf{U} = (\mathbf{H}\mathbf{H}^T + \alpha\mathbf{I})^{-1}\mathbf{H}\mathbf{T}^T \quad (2)$$

By using a gradient descent (Deng and Yu 2011b) algorithm to minimize the the least squares objective in (1) and deriving the gradient of \mathbf{W} in the basic module, we obtain

$$\frac{\partial f_{dsn}}{\partial \mathbf{W}} = 2\mathbf{X} [\mathbf{H}^T \circ (\mathbf{I} - \mathbf{H}^T) \circ (\mathbf{U}\mathbf{U}^T \mathbf{H} - \mathbf{U}\mathbf{T})^T] \quad (3)$$

where \circ denotes element-wise multiplication and \mathbf{I} is the matrix of all ones.

The "convex" solution accentuates the role of convex optimization in learning the output network weights \mathbf{U} in each basic module (Deng and Yu 2011a). Many basic modules are often stacked up with one on top of another to form a deep model. More specifically, the input units of a higher module can include the output units of the lowest module and optionally the raw input feature in the DSN (Deng and Yu 2011b). For obtaining the higher-order information in the data, DSN has recently been extended to Tensor-DSN (T-DSN) (Hutchinson, Deng, and Yu 2013), which's the basic module is to replace a linear map from hidden representation to output with a bilinear mapping. It retains the scalable structure of DSN and provides the higher-order feature interactions missing in DSN.

Sparse Deep Stacking Network

The S-DSN is a sparse case of the DSN. The stacking operation of the S-DSN is the same as that for the DSN described in (Deng and Yu 2011b). The general paradigm is to use the output vector of lower module and the original input vector to form the expanded "input" vector for the higher module of the DSN. The modular architecture of S-DSN is different from that of DSN. We consider the sigmoid function and the ReLU function; and the sparse penalties are added into the hidden units of modular architecture.

Sparse Module

The output of upper-layer is $\mathbf{Y} = \mathbf{U}^T \hat{\mathbf{H}}$ and the hidden layer output is as follow:

$$\hat{\mathbf{H}} = \phi(\mathbf{W}^T \mathbf{X}) \in R^{L \times N} \quad (4)$$

where $\phi(a)$ is the sigmoid activation function $\sigma(a)$ or the ReLU activation function $\max(0, a)$. For simplicity, let $\mathcal{H} = \{1, 2, \dots, L\}$ denote the set of all hidden units. \mathcal{H} is divided into G groups, where G is the number of groups. The g th group is denoted by \mathcal{G}_g , where $\mathcal{H} = \bigcup_{g=1}^G \mathcal{G}_g$ and $\bigcap_{g=1}^G \mathcal{G}_g = \emptyset$. So, $\hat{\mathbf{H}} = [\hat{\mathbf{H}}_{\mathcal{G}_1, :}; \dots; \hat{\mathbf{H}}_{\mathcal{G}_g, :}; \dots; \hat{\mathbf{H}}_{\mathcal{G}_G, :}]$.

The parameters \mathbf{U} and \mathbf{W} are learned to minimize the least squares objective:

$$\min_{\mathbf{U}, \mathbf{W}} f_{sd sn} = \|\mathbf{U}^T \hat{\mathbf{H}} - \mathbf{T}\|_F^2 + \alpha \|\mathbf{U}\|_F^2 + \beta \Psi(\hat{\mathbf{H}}) \quad (5)$$

where α is a regularization parameter of upper-layer weight matrix \mathbf{U} , β is a regularization constant of the activation of the hidden units and $\Psi(\hat{\mathbf{H}})$ represents the imposed penalty over sparse representations $\hat{\mathbf{H}}$. Typically, the l_1 norm is conducted as a penalty to explicitly enforce sparsity on each sparse representation. It is described as:

$$\Psi(\hat{\mathbf{H}}) = \sum_{i=1}^N \|\hat{\mathbf{h}}_i\|_1 \quad (6)$$

where $\hat{\mathbf{h}}_i$ is the representation of i th example ($i = 1, \dots, N$).

In neural networks, sparse representations are advantageous for classification (Ranzato et al. 2007). Moreover, group sparse representations (Bengio et al. 2009) can learn the statistical dependencies between hidden units and lead to better performance (Luo et al. 2011). To implement the dependencies, we averagely divide hidden units into non-overlapping groups to restrain the dependencies within these groups and force hidden units in a group to compete with each other (Luo et al. 2011). Luckily, a mixed-norm regularization (l_1/l_2 -norm) can be conducted in the modular architecture to achieve group sparse representations. Following (Luo et al. 2011), we consider the mixed-norm regularization, which is as follows:

$$\Psi(\hat{\mathbf{H}}) = \sum_{g=1}^G \|\hat{\mathbf{H}}_{\mathcal{G}_g,:}\|_{1,2} \quad (7)$$

where $\hat{\mathbf{H}}_{\mathcal{G}_g,:}$ is the representation matrix associated to those intra-modality data belonging to the g th group and the l_1/l_2 -norm is defined as

$$\|\hat{\mathbf{H}}_{\mathcal{G}_g,:}\|_{1,2} = \sum_{i=1}^N \sqrt{\sum_{j \in \mathcal{G}_g} \hat{h}_{j,i}^2} \quad (8)$$

Learning Weights- Algorithm

Once the lower-layer weight matrix \mathbf{W} are fixed, $\hat{\mathbf{H}}$ are also determined uniquely. Then solving the upper-layer weight matrix \mathbf{U} can be formulated as a convex optimization problem:

$$\min_{\mathbf{U}} f_{sd sn}^u = \|\mathbf{U}^T \hat{\mathbf{H}} - \mathbf{T}\|_F^2 + \alpha \|\mathbf{U}\|_F^2 \quad (9)$$

which has a closed-form solution:

$$\mathbf{U} = (\hat{\mathbf{H}}\hat{\mathbf{H}}^T + \alpha\mathbf{I})^{-1}\hat{\mathbf{H}}\mathbf{T}^T \quad (10)$$

There are two algorithms for learning the lower-layer weight matrix \mathbf{W} . First, given fixed current \mathbf{U} , \mathbf{W} can be optimized using a gradient descent algorithm (Deng and Yu 2011a) to minimize the squared error objective:

$$\min_{\mathbf{W}} f_{sd sn}^1 = \|\mathbf{U}^T \hat{\mathbf{H}} - \mathbf{T}\|_F^2 + \beta \Psi(\hat{\mathbf{H}}) \quad (11)$$

Algorithm 1 Training Algorithm of Sparse Modular

- 1: **Input:** Data matrix \mathbf{X} , label matrix \mathbf{T} , parameters $\theta = \{\epsilon, \alpha, \beta, G\}$ and training epochs E .
- 2: **Initialize:** Projection Matrix \mathbf{W} are initialized with small random values.
- 3: Given \mathbf{W} , calculate $\hat{\mathbf{H}}$ by Eq. (4).
- 4: Update \mathbf{W} by Eq. (16).
- 5: Repeat 3-4 E epochs (or until convergence).
- 6: **Output** weight matrix \mathbf{W} .

and deriving the gradient, we obtain

$$\begin{aligned} \frac{\partial f_{sd sn}^1}{\partial \mathbf{W}} = & 2\mathbf{X} \left[d\phi(\hat{\mathbf{H}}^T) \circ (\mathbf{U}\mathbf{U}^T \hat{\mathbf{H}} - \mathbf{U}\mathbf{T})^T \right] \\ & + 2\beta\mathbf{X} \left[d\phi(\hat{\mathbf{H}}^T) \circ \hat{\mathbf{H}}^T \circ \tilde{\mathbf{H}}^T \right] \end{aligned} \quad (12)$$

where \circ denotes element-wise multiplication, $\circ/$ denotes element-wise division, $\tilde{\mathbf{H}}$ that it's element is $\tilde{h}_{j,i} = \sqrt{\sum_{j \in \mathcal{G}_g} \hat{h}_{j,i}^2}$, $d\phi(\hat{\mathbf{H}}^T)$ denotes element-wise gradient computation, $d\phi(a)$ is the gradient of the activation function. When $\phi(a)$ is the sigmoid activation function, $d\phi(a) = \sigma(a) \times (1 - \sigma(a))$ and when $\phi(a)$ is the ReLU activation function, $d\phi(a)$ is described as:

$$d\phi(a) = \begin{cases} 1, & a > 0; \\ 0, & a \leq 0. \end{cases} \quad (13)$$

To ReLU activation function, we follow the hypothesis (Glorot, Bordes, and Bengio 2011) that the hard non-linearities do not hurt the optimization so long as the gradient can be propagated to many hidden units.

Second, for faster moving \mathbf{W} towards a direction that finds the optimal points, the deterministic nonlinear relationship between \mathbf{U} and \mathbf{W} is used to compute the gradient. By plugging (10) into criterion (5), the least squares objective is rewritten as:

$$\begin{aligned} \min_{\mathbf{W}} f_{sd sn}^2 = & \|[(\hat{\mathbf{H}}\hat{\mathbf{H}}^T + \alpha\mathbf{I})^{-1}\hat{\mathbf{H}}\mathbf{T}^T]^T \hat{\mathbf{H}} - \mathbf{T}\|_F^2 + \\ & \alpha \|(\hat{\mathbf{H}}\hat{\mathbf{H}}^T + \alpha\mathbf{I})^{-1}\hat{\mathbf{H}}\mathbf{T}^T\|_F^2 + \beta \Psi(\hat{\mathbf{H}}) \end{aligned} \quad (14)$$

However, when regularization is used in the objective function (5) (i.e. $\alpha > 0$), the gradient of $f_{sd sn}^2$ can be very complicated. To simplify the gradient we assume $\alpha = 0$ in $f_{sd sn}^2$. So, second term of $f_{sd sn}^2$ is equivalent to zero. Similar to (Deng and Yu 2011b), then we derive the gradient $\frac{\partial f_{sd sn}^2}{\partial \mathbf{W}}$ and obtain

$$\begin{aligned} \frac{\partial f_{sd sn}^2}{\partial \mathbf{W}} = & 2\mathbf{X} \left[d\phi(\hat{\mathbf{H}}^T) \circ [\hat{\mathbf{H}}^\dagger (\hat{\mathbf{H}}\mathbf{T}^T)(\mathbf{T}\hat{\mathbf{H}}^\dagger) - \mathbf{T}^T(\mathbf{T}\hat{\mathbf{H}}^\dagger)] \right] \\ & + 2\beta\mathbf{X} \left[d\phi(\hat{\mathbf{H}}^T) \circ \hat{\mathbf{H}}^T \circ \tilde{\mathbf{H}}^T \right] \end{aligned} \quad (15)$$

where $\hat{\mathbf{H}}^\dagger = \hat{\mathbf{H}}^T (\hat{\mathbf{H}}\hat{\mathbf{H}}^T)^{-1}$ and $d\phi(\cdot)$ and $\tilde{\mathbf{H}}$ are defined in (12).

The algorithm then updates \mathbf{W} using the gradient defined in (12) and (15) as

$$\mathbf{W} = \mathbf{W} - \epsilon \frac{\partial f_{sd sn}^1}{\partial \mathbf{W}} \quad \text{or} \quad \mathbf{W} = \mathbf{W} - \epsilon \frac{\partial f_{sd sn}^2}{\partial \mathbf{W}} \quad (16)$$

Algorithm 2 Training Algorithm of S-DSN

- 1: **Input:** Data \mathbf{X} , label \mathbf{T} , parameters $\theta = \{\epsilon, \alpha, \beta, G\}$, training epochs E and the number of layers K .
 - 2: **Initialize:** $\mathbf{X}^1 = \mathbf{X}$ and $k = 1$.
 - 3: **while** $k \leq K$
 - 4: Given \mathbf{X}^k , \mathbf{T} , θ and E , optimize \mathbf{W}^k by **Algorithm 1**.
 - 5: Given \mathbf{X}^k and \mathbf{W}^k , calculate $\hat{\mathbf{H}}^k$ by Eq. (4), \mathbf{U}^k by Eq. (10) and $\mathbf{Y}^k = (\mathbf{U}^k)^T \hat{\mathbf{H}}^k$.
 - 6: $\mathbf{X}^{k+1} = [\mathbf{X}; \mathbf{Y}^k]$.
 - 7: **end while**
 - 8: **Output** weight matrix $\mathbf{W}^k (k = 1, \dots, K)$.
-

where ϵ is a learning rate. The weight matrix learning process is outlined in **Algorithm 1**.

The S-DSN Architecture

The sparse SNNM module described in the above subsection is used to construct the K -layers S-DSN architecture, where K is the number of layers. In k th sparse module we denote the input by \mathbf{X}^k , hidden representations by $\hat{\mathbf{H}}^k$, output by \mathbf{Y}^k , label matrix by \mathbf{T} and weight matrix by \mathbf{W}^k and \mathbf{U}^k . Given input data \mathbf{X} and label \mathbf{T} , when $k = 1$, $\mathbf{X}^1 = \mathbf{X}$. Then the general paradigm of S-DSN can be decomposed in three phases:

- Step 1: Train the k th sparse module to minimize the least squares error between \mathbf{Y}^k and \mathbf{T} .
- Step 2: Generate the input \mathbf{X}^{k+1} of the $k + 1$ th sparse module by adding the output \mathbf{Y}^k of k th sparse module.
- Step 3: Iterate as in Step 1 and Step 2 to construct the S-DSN architecture.

We summarize the optimization of S-DSN in **Algorithm 2**. For capturing the sparse representation from raw data, this paper proposes the S-DSN, which is implemented by penalizing the hidden unit activations and rectifying the negative of outputs of hidden units activations. Due to the simple structure of each module, the S-DSN still retains the computational advantage of the DSN in parallelism and scalability during learning all parameters.

Experiments

We present experimental results on four databases: the Extended YaleB database, the AR face database, Caltech101 and 15 scene categories.

- Extended YaleB database: this database contains 2,414 frontal face images of 38 people. There are about 64 images for each person. The original images were cropped and normalized to 192×168 pixels.
- AR database: this database consists of over 4,000 color images of 126 people. Each person has 26 face images taken during two sessions. These images include more facial variations, including different illumination conditions, different expressions, and different facial "disguises" (sunglasses and scarves). Following the standard

evaluation procedure, we use a subset of the database consisting of 2,600 images from 50 male subjects and 50 female subjects. Each face image was also cropped and normalized to 165×120 pixels.

- Caltech-101: this database [10] contains 9144 images belonging to 101 classes, with about 40 to 800 images per class. Most images of Caltech-101 are with medium resolution, i.e., about 300×300 .
- 15-Scene: this data set, compiled by several researchers [11, 20, 24], contains a total of 4485 images falling into 15 categories, with the number of images per category ranging from 200 to 400. The categories include living room, bedroom, kitchen, highway, mountain and et al.

According to (Jiang, Lin, and Davis 2013), the four databases are preprocessed ¹: in the Extended YaleB database and AR face database, each face image is projected onto a n -dimensional feature vector with a randomly generated matrix from a zero-mean normal distribution. The dimension of a random-face feature in Extended YaleB is 504 while the dimension in AR face is 540. In face databases the n -dimensional features of each image are normalized to $[-1, 1]$. For the Caltech101 database, we first extract sift descriptors from 16×16 patches, which are densely sampled from each image on a dense grid with 6-pixels stepsize; then we extract the spatial pyramid feature based on the extracted sift features with three grids of size 1×1 , 2×2 and 4×4 . To train the codebook for spatial pyramid, we use the standard k -means clustering with $k = 1024$. For the 15 scene category database, we compute the spatial pyramid feature using a four-level spatial pyramid and a SIFT-descriptor codebook with a size of 200. Finally, the spatial pyramid features are reduced to 3000 dimensions by PCA.

The matrix parameters are initialized with small random values sampled from a normal distribution with zero mean and standard deviation of 0.01. For simplicity, we use the constant learning rate ϵ chosen from $\{20, 15, 5, 2, 1, 0.2, 0.1, 0.05, 0.01, 0.001\}$, the regularization parameter α chosen from $\{1, 0.5, 0.1\}$, the sparse regularization parameter β chosen from $\{0.1, 0.05, 0.01, 0.001, 0.0001\}$ and the group number G chosen from $\{2, 4, 5, 10, 20\}$. In all experiments, we only train 5 epochs, the number of hidden units is 500 and the number of layers is 2. For each data set, each experiment is repeated 10 times with random selected training and testing images, and the average precision is reported. In the rest of this paper, we denote that S-DSN(sigm) indicates S-DSN with sigmoid function; S-DSN(relu) indicates S-DSN with ReLU function; DSN-1, S-DSN(sigm)-1 and S-DSN(relu)-1 respectively indicate one-layer DSN, S-DSN(sigm) and S-DSN(relu); DSN-2, S-DSN(sigm)-2 and S-DSN(relu)-2 respectively indicate two-layer DSN, S-DSN(sigm) and S-DSN(relu).

Table 1: Hoyer’s sparseness measures (HSM) on Extended YaleB and AR databases. We train on 15 (10) samples per category for Extended YaleB (AR) and the rest for testing. For two databases, the number of hidden units is 500, the group sizes for S-DSN(sigm) and S-DSN(relu) are 4 and the number of layers is 2. In Extended YaleB, $\epsilon = 0.2$ and $\alpha = 0.5$ are used for DSN; $\epsilon = 0.2$, $\alpha = 0.5$ and $\beta = 0.001$ are used for S-DSN(sigm). In Extended YaleB $\epsilon = 0.05$ and $\alpha = 1$ are used for DSN; $\epsilon = 0.05$, $\alpha = 1$ and $\beta = 0.0001$ are used for S-DSN(relu).

	layers	S-DSN(sigm)		DSN	
		HSM	Acc. (%)	HSM	Acc. (%)
Extended YaleB	1	0.096	91.4	0.010	88.9
	2	0.111	92.0	0.012	89.4
	layers	S-DSN(relu)		DSN	
		HSM	Acc. (%)	HSM	Acc. (%)
AR	1	0.286	93.2	0.003	80.2
	2	0.306	93.5	0.003	81.2

Table 2: Recognition Results Using Random-Face Features on the Extended YaleB Database

Methods	Acc. (%)	Methods	Acc. (%)
SRC	97.2	LC-KSVD	96.7
DSN-1	96.6	DSN-2	96.9
S-DSN(sigm)-1	96.9	S-DSN(sigm)-2	97.4
S-DSN(relu)-1	96.1	S-DSN(relu)-2	96.7

Sparseness Comparisons

Before presenting classification results, we first show the sparseness of S-DSN(sigm) and S-DSN(relu) compared to DSN. We use Hoyer’s sparseness measure (HSM) (Hoyer 2004) to figure out how sparse representations learned by the S-DSN(sigm), S-DSN(relu) and DSN. This measure has good properties, which is in the interval $[0, 1]$ and on a normalized scale. Its value more close to 1 means that there are more zero components in the vector. We perform comparisons on Extended YaleB and AR databases, and results are reported in Table 1. The sparseness results show that S-DSN(sigm) and S-DSN(relu) have higher sparseness and higher recognition accuracy. Table 1 compares the network HSM of the S-DSN(sigm) and the S-DSN(relu) to that of DSN. We observe that the average sparseness of two layers S-DSN(sigm) is about 0.105 (Extended YaleB) and the average sparseness of two layers S-DSN(relu) is about 0.291 (AR). In contract, the average sparseness of two layers DSN is on average below 0.02 in the databases. It can be seen that the S-DSN can learn sparser representations. Due to space reasons, Figure 1 only visualizes the activation probabilities of first hidden layer, which are computed under the S-DSN(relu) and the DSN given an image from test set of AR.

Results

Face Recognition Extended YaleB: We randomly select half (32) of the images per category as training and the other half for testing. The parameters are selected as follow: in DSN

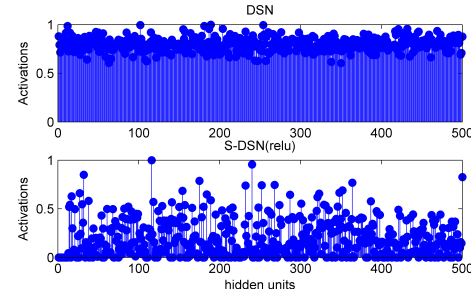


Figure 1: Activation probabilities of first hidden layer are computed under DSN and S-DSN(relu) on the AR database. Activation probabilities be normalized by dividing the maximum of activation probabilities.

Table 3: Recognition Results Using Random Face Features on the AR Face Database

Methods	Acc. (%)	Methods	Acc. (%)
SRC	97.5	LC-KSVD	97.8
DSN-1	97.6	DSN-2	97.8
S-DSN(sigm)-1	97.9	S-DSN(sigm)-2	98.1
S-DSN(relu)-1	97.6	S-DSN(relu)-2	97.8

$\epsilon = 0.1$ and $\alpha = 0.5$; in S-DSN(sigm) $\epsilon = 0.1$, $\alpha = 0.5$, $G = 2$, and $\beta = 0.01$; in S-DSN(relu) $\epsilon = 0.01$, $\alpha = 2$, $G = 5$, and $\beta = 0.001$. AR: For each person, we randomly select 20 images for training and the other 6 for testing. In our experiments, $\epsilon = 0.1$ and $\alpha = 0.5$ are used in DSN; $\epsilon = 0.1$, $\alpha = 0.5$, $G = 4$, and $\beta = 0.001$ are used in S-DSN(sigm); $\epsilon = 0.01$, $\alpha = 1$, $G = 4$, and $\beta = 0.001$ are used in S-DSN(relu).

We compare S-DSN with DSN (Deng and Yu 2011b), and LC-KSVD (Jiang, Lin, and Davis 2013) and SRC (Wright et al. 2009) algorithms, which reported state-of-the-art results on those two databases. The experimental results are summarized in Table 2 and Table 3, respectively. S-DSN(sigm) achieves better results than DSN, LC-KSVD and SRC. From Table 2 S-DSN(sigm)-1 is better than LC-KSVD and has about 0.2% improvement in Extended YaleB. From Table 3, S-DSN(sigm)-1 and S-DSN(sigm)-2 are also better than LC-KSVD and have about 0.1% and 0.3% improvement in AR, respectively. In addition, we compare with LC-KSVD in terms of the computation time for classifying one test image. S-DSN has a faster inference because it can directly learn projection dictionaries. As shown in Table 4, S-DSN is 7 times faster than LC-KSVD.

15 Scene Category: Following the common experimental settings, we randomly choose 100 images from each class for training data and the rest for test data. In our experiments, $\epsilon = 20$ and $\alpha = 0.1$ are used in DSN; $\epsilon = 20$, $\alpha = 0.1$, $G = 4$, and $\beta = 0.05$ are used in S-DSN(sigm); $\epsilon = 15$, $\alpha = 0.1$, $G = 4$, and $\beta = 0.001$ are used in S-DSN(relu).

We compare our results with SRC (Wright et al. 2009),

Table 4: Inference Time (ms) for a Test Image on the Extended YaleB Database

Methods	SRC	LC-KSVD	S-DSN(relu)
Average time	20.121	0.502	0.069

¹they can be downloaded from: <http://www.umiacs.umd.edu/~zhuolin/projectlcksvd.html>

Table 5: Recognition Results Using Spatial Pyramid Features on the 15 Scene Category Database

Methods	Acc. (%)	Methods	Acc. (%)
ITDL	81.1	ISPR+IFV	91.1
SR-LSR	85.7	ScSPM	80.3
LLC	89.2	SRC	91.8
LC-KSVD	92.9	DeepSC	83.8
DeCAF	88.0	DSFL+DeCAF	92.8
DSN-1	96.7	DSN-2	97.0
S-DSN(sigm)-1	96.5	S-DSN(sigm)-2	97.1
S-DSN(relu)-1	98.8	S-DSN(relu)-2	98.8

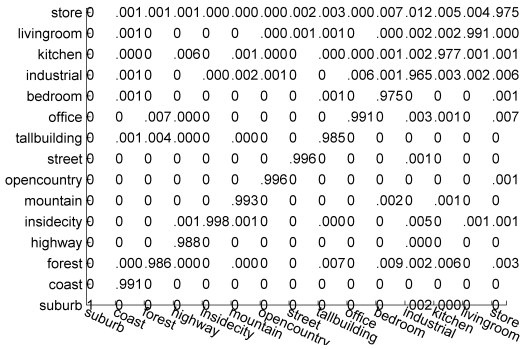


Figure 2: The confusion matrix on the 15 scene category database.

LC-KSVD (Jiang, Lin, and Davis 2013), DeepSC (He et al. 2014), DSN (Deng and Yu 2011b) and other state-of-the-art approaches: ScSPM (Yang et al. 2009), LLC (Wang et al. 2010), ITDL (Qiu, Patel, and Chellappa 2014), ISPR+IFV (Lin et al. 2014), SR-LSR (Li and Guo 2014), DeCAF (Donahue et al. 2014), DSFL+DeCAF (Zuo et al. 2014). The detailed comparison results are shown in Table 5. Compared to LC-KSVD, S-DSN(relu)-1’s performance is much better, since it makes a 5.9% improvement. It also registers about 1.8% improvement over the deep models: DeepSC, DeCAF, DSFL+DeCAF and DSN. As Table 5 shows, we see that S-DSN(relu) performs best among all existing methods. The confusion matrix for the S-DSN(relu) are further shown in Figure 2, from which we can see that the misclassification errors of industrial and store are higher than others.

Caltech101: Following the common experimental settings, we train on 5, 10, 15, 20, 25, and 30 samples per category and test on the rest. Due to space reasons, we only give the parameters for 30 training samples per category: $\epsilon = 0.2$ and $\alpha = 0.5$ are used in DSN; $\epsilon = 0.2$, $\alpha = 0.5$, $G = 4$, and $\beta = 0.01$ are used in S-DSN(sigm); $\epsilon = 0.05$, $\alpha = 0.5$, $G = 2$, and $\beta = 0.001$ are used in S-DSN(relu).

We evaluate our approach using spatial pyramid features and compare with SRC (Wright et al. 2009), LC-KSVD (Jiang, Lin, and Davis 2013), DeepSC (He et al. 2014), DSN (Deng and Yu 2011b) and other approaches ScSPM (Yang et al. 2009), LLC (Wang et al. 2010), LRSC (Zhang et al. 2013), LCLR (Jiang, Guo, and Peng 2014). The average classification rates are reported in Table 6. From these results, S-DSN(relu)-1 outperforms the other competing dictionary learning approaches, including LC-KSVD, LRSC, and SRC; and has 1.6% improvement. S-DSN(relu) also reg-

Table 6: Recognition Results Using Spatial Pyramid Features on the Caltech101 Database

Methods	5	10	15	20	25	30
ScSPM	-	-	67.0	-	-	73.2
SRC	48.8	60.1	64.9	67.7	69.2	70.7
LLC	51.2	59.8	65.4	67.7	70.2	73.4
LC-KSVD	54.0	63.1	67.7	70.5	72.3	73.6
LRSC	55.0	63.5	67.1	70.3	72.7	74.4
LCLR	53.4	62.8	67.2	70.8	72.9	74.7
DSN-1	53.6	61.8	67.7	70.2	72.0	74.6
DSN-2	54.9	63.4	68.2	70.5	72.9	74.7
S-DSN(sigm)-1	54.0	62.3	67.6	70.2	72.1	74.7
S-DSN(sigm)-2	55.4	63.7	68.3	70.8	73.2	74.9
S-DSN(relu)-1	55.4	63.8	68.7	71.2	73.5	76.0
S-DSN(relu)-2	55.6	64.2	69.0	71.3	73.6	76.2

isters about 1.5% improvement over a deep model: DSN. Note that 76.5% accuracy achieved by our method (the number of hidden units is 1000) is also competitive with the 78.4% reported in DeepSC.

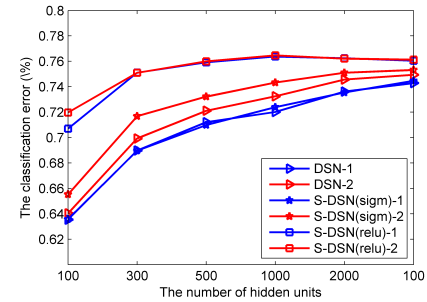


Figure 3: Effect of the number of hidden units used in S-DSN(sigm), S-DSN(relu) and DSN on recognition accuracy.

We examine how performance of the proposed S-DSN changes when varying the number of hidden units. We randomly select 30 images per category for training data and the rest for test data. We consider six settings where the number of hidden units changes from 100 to 3000 and compare the results with DSN. As reported the results in Figure 3, our approaches maintain high classification accuracies and outperform the DSN model. When increasing the number of hidden units, the accuracy of the system improves for S-DSN(sigm), S-DSN(relu) and DSN.

Effects of Number of Layers: The deep framework utilizes multiple-layers of feature abstraction to get a better representation for images. From Tables 2, 3, 5 and 6, we check the effect of varying the number of layers and the classification accuracy improves as the number of layers increases. In addition, compared to the dictionary learning approaches, S-DSN has a faster inference and a deep architecture. Moreover, S-DSN has a good performance in image classification.

Conclusion

In this paper, we present an improved DSN model, S-DSN, for image classification. S-DSN is constructed by stacking many sparse SNNM modules. In each sparse SNNM module, the lower-layer weights and the upper-layer weights are solved by using the convex optimization and the gradient

descent algorithm. We use the S-DSN to further extract the sparse representations from the random face features and spatial pyramid features for image classification. Experimental results show that S-DSN yields very good classification results on four public databases with only a linear classifier.

Acknowledgments

This work was partially supported by the National Science Fund for Distinguished Young Scholars under Grant Nos. 61125305, 61472187, 61233011 and 61373063, the Key Project of Chinese Ministry of Education under Grant No. 313030, the 973 Program No. 2014CB349303, Fundamental Research Funds for the Central Universities No. 30920140121005, and Program for Changjiang Scholars and Innovative Research Team in University No. IRT13072.

References

- [Bengio et al. 2009] Bengio, S.; Pereira, F.; Singer, Y.; and Strelow, D. 2009. Group sparse coding. In *Proceedings of the Neural Info. Processing Systems*, 399–406.
- [Deng and Yu 2011a] Deng, L., and Yu, D. 2011a. Accelerated parallelizable neural network learning algorithm for speech recognition. In *Proceedings of the Interspeech*, 2281–2284.
- [Deng and Yu 2011b] Deng, L., and Yu, D. 2011b. Deep convex networks: A scalable architecture for speech pattern classification. In *Proceedings of the Interspeech*, 2285–2288.
- [Deng and Yu 2013] Deng, L., and Yu, D. 2013. Deep learning for signal and information processing. *Foundations and Trends in Signal Processing* 2-3:197–387.
- [Deng, He, and Gao 2013] Deng, L.; He, X.; and Gao, J. 2013. Deep stacking networks for information retrieval. In *Proceedings of IEEE Conf. on Acoustics, Speech, and Signal Processing*, 3153–3157.
- [Deng, Yu, and Platt 2012] Deng, L.; Yu, D.; and Platt, J. 2012. Scalable stacking and learning for building deep architectures. In *Proceedings of IEEE Conf. on Acoustics, Speech, and Signal Processing*, 2133–2136.
- [Donahue et al. 2014] Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the Int'l Conf. Machine Learning*, 647–655.
- [Glorot and Bengio 2010] Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Int'l Conf. Artificial Intelligence and Statistics*, 249–256.
- [Glorot, Bordes, and Bengio 2011] Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep sparse rectifier neural networks. In *Proceedings of the Int'l Conf. Artificial Intelligence and Statistics*, 315–323.
- [Gregor and LeCun 2010] Gregor, K., and LeCun, Y. 2010. Learning fast approximations of sparse coding. In *Proceedings of the Int'l Conf. Machine Learning*, 399–406.
- [He et al. 2014] He, Y.; Kavukcuoglu, K.; Wang, Y.; Szlam, A.; and Qi, Y. 2014. Unsupervised feature learning by deep sparse coding. In *Proceedings of SIAM Int'l Conf. on Data Mining*, 902–910.
- [Hoyer 2004] Hoyer, P. 2004. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research* 5:1457–1469.
- [Huang et al. 2013] Huang, J.; Nie, F.; Huang, H.; and Ding, C. 2013. Supervised and projected sparse coding for image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 438–444.
- [Hutchinson, Deng, and Yu 2013] Hutchinson, B.; Deng, L.; and Yu, D. 2013. Tensor deep stacking networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 35:1944–1957.
- [Jiang, Guo, and Peng 2014] Jiang, Z.; Guo, P.; and Peng, L. 2014. Locality-constrained low-rank coding for image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2780–2786.
- [Jiang, Lin, and Davis 2013] Jiang, Z.; Lin, Z.; and Davis, L. 2013. Label consistent k-svd learning a discriminative dictionary for recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 35:2651–2664.
- [Lee et al. 2007] Lee, H.; Battle, A.; Raina, R.; and Ng, A. 2007. Efficient sparse coding algorithms. In *Proceedings of the Neural Info. Processing Systems*, 801–808.
- [Lee, Ekanadham, and Ng 2008] Lee, H.; Ekanadham, C.; and Ng, A. 2008. Sparse deep belief net model for visual area v2. In *Proceedings of the Neural Info. Processing Systems*, 873–880.
- [Lennie 2003] Lennie, P. 2003. The cost of cortical computation. *Current Biology* 13:493–497.
- [Li and Guo 2014] Li, X., and Guo, Y. 2014. Latent semantic representation learning for scene classification. In *Proceedings of the Int'l Conf. Machine Learning*, 532–540.
- [Lin et al. 2014] Lin, D.; Lu, C.; Liao, R.; and Jia, J. 2014. Learning important spatial pooling regions for scene classification. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, 3726–3733.
- [Luo et al. 2011] Luo, H.; Shen, R.; Niu, C.; and Ullrich, C. 2011. Sparse group restricted boltzmann machines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 429–433.
- [Nair and Hinton 2010] Nair, V., and Hinton, G. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the Int'l Conf. Machine Learning*, 807–814.
- [Qiu, Patel, and Chellappa 2014] Qiu, Q.; Patel, V.; and Chellappa, R. 2014. Information-theoretic dictionary learning for image classification. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 36:2173–2184.
- [Ranzato et al. 2007] Ranzato, M.; Boureau, Y.; Chopra, S.; and LeCun, Y. 2007. Efficient learning of sparse representations with an energy-based model. In *Proceedings of the Neural Info. Processing Systems*, 1137–1144.
- [Ranzato, Boureau, and LeCun 2008] Ranzato, M.; Boureau, Y.; and LeCun, Y. 2008. Sparse feature learning for deep belief networks. In *Proceedings of the Neural Info. Processing Systems*, 1185–1192.
- [Wang et al. 2010] Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; and Gong, Y. 2010. Locality-constrained linear coding for image classification. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, 3360–3367.
- [Wolpert 1992] Wolpert, D. 1992. Stacked generalization. *Neural Networks* 5:241–259.
- [Wright et al. 2009] Wright, J.; Yang, M.; Ganesh, A.; Sastry, S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 31:210–227.
- [Yang et al. 2009] Yang, J.; Yu, K.; Gong, Y.; and Huang, T. 2009. Linear spatial pyramid matching using sparse coding for image

classification. In *Proceedings of IEEE Conf. Computer Vision and Pattern Recognition*, 1794–1801.

[Yang et al. 2012] Yang, J.; Zhang, L.; Xu, Y.; and Yang, J. 2012. Beyond sparsity: The role of l_1 -optimizer in pattern classification. *Pattern Recognition* 45:1104–1118.

[Zhang et al. 2013] Zhang, T.; Ghanem, B.; Liu, S.; Xu, C.; and Ahuja, N. 2013. Low-rank sparse coding for image classification. In *Proceedings of the IEEE Int’l Conf. Computer Vision*, 281–288.

[Zhuang et al. 2013] Zhuang, Y.; Wang, Y.; Wu, F.; Zhang, Y.; and Lu, W. 2013. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1070–1076.

[Zuo et al. 2014] Zuo, Z.; Wang, G.; Shuai, B.; Zhao, L.; Yang, Q.; and Jiang, X. 2014. Learning discriminative and shareable features for scene classification. In *Proceedings of the Euro. Conf. Computer Vision*, 552–568.